An Interoperable Data Architecture for Data Exchange in a Biomedical Research Network

Daniel Crichton
NASA Jet Propulsion Laboratory
California Institute of Technology
Dan.Crichton@jpl.nasa.gov

Heather Kincaid
Fred Hutchinson Cancer Research Center
hkincaid@fhcrc.org

Gregory J. Downing
National Institutes of Health
downingg@od.nih.gov

Sudhir Srivastava
National Cancer Institute
srivasts@dcpcepn.nci.nih.gov

J. Steven Hughes
NASA Jet Propulsion Laboratory
California Institute of Technology
Steven.Hughes@jpl.nasa.gov

Abstract

Knowledge discovery and data correlation require a unified approach to basic data management. However, achieving such an approach is nearly impossible with hundreds of disparate data sources, legacy systems, and data formats. This problem is pervasive in the biomedical research community where data models, taxonomies, and data management systems are locally implemented. These local implementations create an environment where interoperability and collaboration between researchers and research institutions are limited. Investigators from this paper demonstrate how technology developed by NASA's Jet Propulsion Laboratory (JPL) for space science can be used to build an interoperable data architecture for bioinformatics. JPL has taken a novel approach towards solving this problem by exploiting web technologies usually dedicated to e-commerce, combined with a rich, metadata-based environment. This paper discusses the approach taken to develop a prototype data architecture for the discovery and validation of disease biomarkers within a biomedical research network. Biomarkers are measured parameters of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention. Biomarkers are of growing importance in the biomedical research for therapeutic discovery, disease prevention, and detection. A bioinformatics infrastructure is crucial to support the integration and analysis of large, complex biological and epidemiologic datasets.

1. Introduction

The Early Detection Research Network (EDRN), supported by the National Cancer Institute (NCI), is a consortium of investigators focusing on the research, development, evaluation, and validation of new tests to support the early detection of cancer [4]. The network consists of 18

Biomarkers Development Laboratories (BDL), 3 Biomarkers Validation Laboratories (BVL), 8 Clinical Epidemiological Centers (CEC), and a Data Management and Coordination Center (DMCC) distributed across the United States. It is unique in its capability of advancing translational research of molecular, genetic, and other biomarkers in human cancer detection and risk assessment.

A principal goal of the EDRN consortium is to provide collaboration between each of the EDRN centers through a knowledge environment. The knowledge environment allows principal investigators and researchers the opportunity to share scientific research with an ultimate goal of interconnecting each of the centers in order to establish an informatics enterprise supporting biomedical research in a multi-database, multi-disciplinary environment.

In order to accomplish this task, a pilot project was initiated to study the feasibility of inter-connecting laboratory databases at the centers while retaining their geographic and heterogeneous implementations. The pilot identified four specific goals for the system that included 1) understanding where the data resources reside, 2) understanding how data is accessed in each system, 3) interpreting and defining the underlying data models for each system, and 4) building a ubiquitous interface that demonstrates enterprise-wide query capabilities. A pilot project team was formed consisting of expertise from the National Institutes of Health (NIH), the National Cancer Institute (NCI), the Jet Propulsion Laboratory (JPL), the EDRN DMCC, and investigators based at major academic research universities. The team focused on three key areas that include policy, data engineering and system architecture. Each of these areas required specific expertise in the domain to overcome the challenges inherent with deploying a data architecture to enable interoperability across the EDRN sites.

2. Project scope and planning

Project planning began with a user assessment study of the network's bioinformatics needs to support biomarker discovery and validation. Researchers identified data and information sharing needs as a high priority, specifically, the capability to identify biological samples stored in various network laboratories, the ability to link information about biospecimens to descriptive epidemiology characteristics of human research subjects, and the ability to transfer large and complex data for statistical analysis. They also wanted the ability to share knowledge and to identify work from other centers to explore future collaborative partners. They preferred to work in a web-based environment and to continue using their own legacy information systems. Researchers identified longer term needs for building a biomarkers knowledge environment that included providing analytical tools for genomic, proteomic, and imaging data.

Expertise was retained from JPL to explore integration of software components developed for space science [1] to deploy a data management architecture for the EDRN. The goal was to build a software system that supported the policy and data engineering requirements necessary to build a knowledge system. JPL used the Planetary Data System (PDS) [9], a geographically distributed archive system for NASA's planetary missions, as a model to share experiences with the EDRN on how to build and support a federated, heterogeneous science data system.

The DMCC serves as the central coordinating mechanism for the EDRN. Specifically, it focuses on providing technology for central communication and sharing collaborative data and basic network-wide informatics support, network coordination and logistic support including an institutional review board process for human subjects protection issues and support for statistical and computational methods for data analysis. They also coordinated the pilot project and worked with each of the centers in order to facilitate its progress.

One of the key areas of focus for this project was on generating and managing metadata. Metadata has proven to be a key in the ability to interoperate heterogeneous data systems. A great deal of emphasis was placed on data engineering to understand how to interpret and interrelate the local data dictionaries for each of the centers. The DMCC had already established an effort collaborating with NCI to develop common data elements (CDEs) describing specimens and human subject populations across the network. The NCI CDE project [11] is designed to standardize and simplify the collection and reporting of data for clincial trials, patient surveys, and cancer patient care by collaboratively developing uniform and explicit data elements based on recognized standards. The CDEs represent a subset of data that are considered by EDRN investigators to be the most important data that will be shared among the network. The core does not represent all the data that might be collected in any given EDRN study. Among many things, the CDEs will promote data sharing and data analysis through the use of common terminology and common data values. This allowed the team to work on mapping the local data dictionaries to the CDEs defined for the network.

The pilot team developed a multi-phase plan to address the data sharing needs of the investigators. The plan addressed two key phases. The first phase was a demonstration that would implement a data architecture with all data housed in separate databases at JPL, and the second was to connect directly to the data hosted at each center. The team identified several centers in the network with existing in-house databases describing biospecimens and currently existing study populations as key targets of the pilot. At the time of this writing, the team has successfully demonstrated the data architecture for the first phase and is developing the interfaces directly with the centers for the second phase.

The working group identified the problem to address in the pilot as biospecimen data that is geographically distributed across heterogeneous databases making the location, retrieval, and use of the data difficult. The objective was to implement a software framework with sufficient layers of abstraction to insulate users from data system specifics while promoting data management best practices at the data systems level. If successfully implemented, this pilot project would serve as a model for building additional research modules for the location and retrieval of research data.

3. Data architecture

The architecture implemented focused on the deployment of a middleware component framework with the objective of interconnecting distributed heterogeneous data systems without having to re-implement the underlying data or object models for these systems. The key architectural objectives included 1) requiring that individual data systems be encapsulated to hide uniqueness, 2) requiring that communication between distributed services use metadata for data interchange, 3) defining a standard data dictionary based on metadata for describing data resources, 4) providing a solution that is both scalable and extensible, 5) providing a standard mechanism for exchanging data system product results across distributed services, and 6) allowing systems using different data dictionaries to be integrated.

Software components from a framework developed by JPL for interconnecting space science data systems called Object Oriented Data Technology (OODT) [2] were used. These components include a query service component for managing distributed queries, a profile service component for managing metadata repositories, and a product service component for integrating with distributed data sources. These components have been implemented using Java, the Extensible Markup Language (XML) [7], and the Common Object Request Broker Architecture (CORBA) [6]. Profile and product servers were instantiated at multiple centers

allowing for the system to scale. Each node of the architecture was capable of exchanging data based on metadata definitions. Figure 1 describes the interaction among each of the servers.

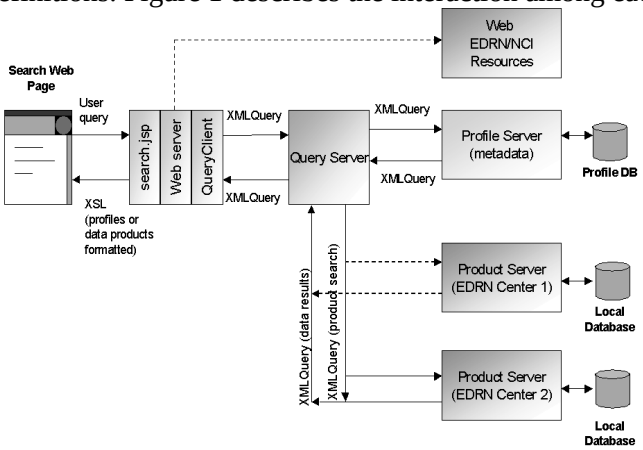

Figure 1. EDRN knowledge system query flow

The component architecture lends itself to a distributed object implementation. Each component of the architecture used XML over CORBA for communicating with other components over the Internet. This is significant since one of the critical requirements of this architecture is to provide interoperability solutions without having to change the implementation of each data system. The architecture accomplishes this by encapsulating each of the individual data systems and focusing communication on standard metadata definitions implemented with the XML specification language.

The profile server manages profiles—sets of resource definitions [1]—about distributed data systems and their products. A profile is a metadata description of the resources known by a node in the distributed framework. These resources are interfaces, data products, or profile servers available in the integrated enterprise. Profiles may be grouped and served by more than one profile server. The query component ties this architecture together by providing and managing the traversal of the integrated digraph node architecture. It also interprets profile definitions that provide mappings between data system nomenclature. The query component also provides the facility to manage concurrent queries across multiple servers to improve performance.

The product server provides the translation necessary to map a product retrieved from a data-system–dependent environment into a neutral format suitable for exchange between systems. The product server architecture is similar to the profile architecture by providing a distributed approach allowing one or more instantiations of product servers across a distributed enterprise. A specific goal of this architecture is to allow heterogeneous data systems to be easily added without changing the way their data is stored.

The Common Object Request Broker Architecture (CORBA) [6] was used along with Java to support a distributed object implementation. For this implementation, one profile server was used to reference each of the product servers for the center's databases. Profile and product server instantiations are uniquely identified by name. These names are used as part of the metadata header encoded to identify enterprise services that can support queries for distributed products. CORBA manages the mapping of a distributed object to a physical location of a profile or product server. As the pilot project progresses, this will allow us to meet our second phase objective to connect directly to the center merely by changing the physical mappings of where the data resides.

XML was chosen since it provides a rich environment for defining and managing metadata. In addition, it was used as an interface specification within CORBA between the nodes of the

system. The interface specification expresses a query within the system. The query definition is implemented independent of any one database, functional, or programming language and is intended to provide an abstract view of both the query expression and the results. The query definition allows for each data system to be encapsulated. This allows various implementations, ranging from the use of relational and object database management systems to the use of flat file and home-grown databases for cataloging and storing data products to exchange information by plugging into a generic query definition.

One of the goals of this architecture is to provide a standard application program interface (API) that will allow for generic science analysis tools to be written that can plug into the architecture to retrieve and correlate data from multiple data sources. For this pilot, a web-based interface was developed using Java Server Pages (JSP). The interface served as a client of the data architecture and allowed for researchers to query distributed databases from a single point.

The data architecture focuses on providing a knowledge environment for supporting interoperability and integration across the EDRN sites. It supports interoperability by building components that support the exchange and management of metadata for describing system resources. It also allows data systems to retain their unique attributes, yet plug into a collaborative enterprise data network allowing for exchange of data content using XML.

It was discovered that architectural goals for space science and biomedical research were very similar, and in fact, the components developed for the space science could be directly infused into the EDRN knowledge environment. By focusing on a framework for supporting basic system interoperability, the architecture was able to provide solutions that not only solve problems within a single discipline, but also support integration of cross-disciplinary databases.

4. Metadata management

Metadata, or data about data, is used to describe both data and non-data resources in the data systems participating in the pilot project. These resource descriptions, referred to in the implementation as “profiles”, are generated based on data system data dictionaries and data models. They focus on the data elements and attributes that characterize the data being managed.

Prior to this pilot project, the DMCC had already begun developing common data elements (CDEs) for core epidemiological and specimen terms across the network, as previously mentioned. The DMCC led the development and implementation of the data elements in collaboration with EDRN investigators at other sites, who represent a variety of clinical and basic science disciplines, NCI, and expertise from other external representatives.

Once the CDEs were well defined, JPL teamed up with investigators from the EDRN to create resource profiles representing resources within the EDRN. The profiles used the data elements to map data resources within the EDRN using a standard vocabulary. For the pilot, many of the data resources provided by the centers existed prior to the creation of the EDRN. Therefore, the data attributes at each center had to be mapped to the EDRN CDEs.

The resource profile was designed with three sections: the profile attributes, the resource attributes, and the profile elements. The “profile attributes” section simply describes the profile itself and contains system level attributes such as profile identifier, type, and status.

The “resource attribute” section generically describes the resource. For this section, the Dublin Core metadata element set for describing electronic resources on the Internet has been adopted. These include attributes such as title, description, and creator. Three additional

resource attributes have been designed to identify the local domain, the resource classification, and the resource's location.

Finally, the "profile element" section provides a non-generic description of the resource by encoding domain specific attributes of the resource and the data that it manages. The elements allow for attributes existing in the local data dictionary to be managed and related to the resources. These attributes are typically extracted from the participating data system's data dictionaries. For example, during the EDRN phase one implementation, a resource that provided gender specific data used "gender" as an attribute name and "2" as a value that represented "female". The term "gender" had also been adopted as a CDE. Another participating resource used "sex" with the value "female". Both attribute/values were encoded into the profiles, and "sex" was designated a synonym enabling interoperability between the two resources. The profile element section uses meta-attributes derived from ISO/IEC 11179 – Specification and Standardization of Data Elements [10].

Once profiles are identified, they are loaded into registries. These registries are searched and allow any system component to identify and retrieve profiles that meet certain constraints. For example, the query service can identify resources that resolve a user query by extracting the query's constraints and searching the profile registry. The query service previously mentioned as part of the OODT software allows for user queries to be sent to all registries in order to locate the appropriate resources.

Creating profiles of distributed resources is an important step to locating resources in individual data systems and understanding how the unique name spaces within each system can be interoperated. JPL's experience in developing and enforcing metadata standards for space science has proven that metadata is a critical precursor to providing interoperability and system federation in a multi-database, multi-institutional environment.

5. Connecting heterogeneous, distributed databases

Federating data systems together is dependent on satisfying four key goals we previously mentioned including 1) understanding where data resources within information systems reside, 2) understanding how data is accessed, 3) interpreting and defining the underlying models that drive these systems, and 4) building cohesive interfaces that satisfy and demonstrate enterprise-wide user queries and scenarios. A key part of the solution implemented is providing data and system abstraction and transparency so the users are presented with a unified view of EDRN. The prototype accomplished this by creating a unified interface that allows scientists to query epidemiological and biospecimen data across multiple databases. In order to integrate databases across the EDRN in this prototype, the resource profile registries previously mentioned were used to manage relationships between the common data elements, the local data dictionaries, and the references to data products residing within a center's database. This allowed the first objective to be satisfied, which is identifying the correct resources within the each site's local database.

Next, product servers, as defined in section 3, were developed for each center which received and translated an XML query expression into a local query for the center database. For the pilot project, this required that each product server be implemented to map the generic metadata model into a local model defined by the center's local data dictionary. The product server also managed the mapping of the data elements identified in the center's resource profile to specific attributes in the data system. The implementation of the product servers met our second and third objectives of being able to access data from individual center databases, as well as mapping the local data dictionaries to a common data dictionary based on the CDEs.

Finally, a web interface was developed using Java Server Pages that plugged into the Java framework and allowed researchers to perform queries of the multiple specimen banks over the Internet as demonstrated in Figure 2. The web interface formulated queries using the XML query expression language supported by the OODT software as previously mentioned. The web interface performed two queries. The first identified those candidate data systems that had a product server that could handle the query, and the second forwarded concurrent queries to each product server that could support the query. Results from each product server were then integrated and displayed. The initial scenario that was identified was then tested against the implemented data sharing system. As an example, a scenario that would return all cancer specimens for men age seventy years or older diagnosed with prostate cancer in the last five years was identified as a target scenario to be solved. This allowed us to meet our fourth objective of developing an interface that supported a previously identified use case scenarios and integrated data from multiple centers in a cohesive interface.

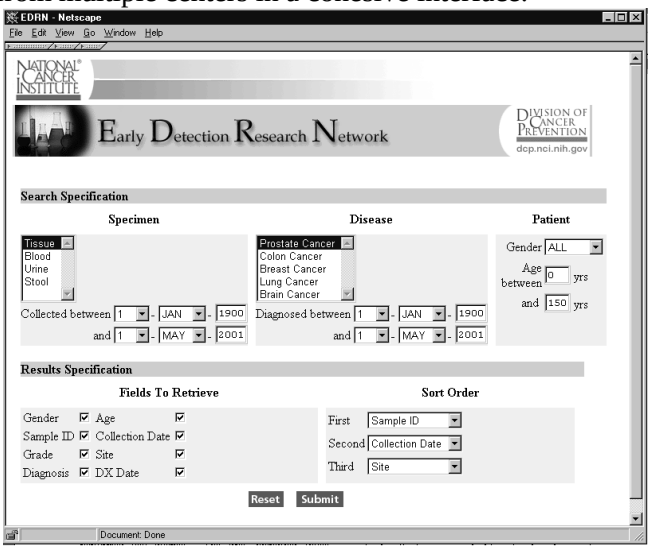

Figure 2. EDRN prototype user interface

There were several challenges encountered in interconnecting these data systems. The first was that the data collected at each center differed due to data acquisition prior to the establishment of the EDRN. Attributes in one database, for example, were missing from another. Also, certain concepts were implied in one database, and explicit in another. These inherent differences made understanding the models essential for interoperability. In addition, the centers used different vendor technologies for implementing their systems. This meant having to develop techniques that neutralized the vendor implementations so that they were compatible. Finally, the institutional access issue was identified as a key challenge and broken into two phases. The first phase was to extract, de-identify confidential attributes, and rebuild the databases at JPL. The second phase, which is currently ongoing, is to connect directly to the databases residing at each center. This allows for the data engineering and system architecture to be solidified independent of the policy issues related to institutional access to data residing at each center.

6. Conclusions

The achievements described here reflect the first milestones of a pilot data architecture that shows the successful deployment of JPL's OODT software to meet the needs of biomedical researchers. The current phase of the project is working to deploy the architecture across

several institutions and is focusing on additional query capabilities and long term plans for integrating analytical tools for the investigators. Future phases will target the creation of data profiles that include metatdata descriptions of biospecimens, biomarkers including assay sensitivity and specificity, research protocols, and publications that ultimately will yield a biomarker knowledge environment. The prototype overcame several inherent technical challenges presented by the heterogeneity of the implementations between the institutions. In addition, the project encountered several administrative and policy challenges. Because of the extensive involvement of human subjects research data, particular attention was applied to network and computer security, privacy and confidentiality considerations, institutional review board policies, and intellectual property associated with data sharing. It is concluded that the software technology transferred from space science disciplines can also be applied to biomedical research. In addition, the experience has shown that the development of metadata is paramount to the successful implementation of a data architecture for biomedical research. Finally, the federation of heterogeneous, disparate databases and data dictionaries have the capability to enhance the archiving, retrieval, and analysis of key biomedical research data and will continue to be further evaluated in additional scenarios and disciplines to address complex imaging, genomic, and proteomic datasets.

7. References

[1] D.J. Crichton, J.S. Hughes, J.J. Hyon, S.C. Kelly, A Distributed Component Framework for Science Data Product Interoperability, The 17th International Conference on Scientific and Technical Data. October 2001. http://oodt.jpl.nasa.gov/doc/papers/italy_codata/italy_paper.pdf

[2] D.J. Crichton, J.S. Hughes, J.J. Hyon, S.C. Kelly, Science Search and Retrieval using XML, The Second National Conference on Scientific and Technical Data, U.S. National Committee for CODATA, National Research Council, March 13-14, 2000, http://oodt.jpl.nasa.gov/doc/papers/codata/paper.pdf.

[3] Biomarkers Knowledge System. http://www1.od.nih.gov/osp/ospp/biomarkers/Biomarkers_Knowledge_System.pdf

[4] Early Detection Research Network. http://edrn.nci.nih.gov/

[5] J.S. Hughes, D.J. Crichton, J.J. Hyon, S.C. Kelly, A Multi-Discipline Metadata Registry for Science Interoperability, Open Forum on Metadata Registries, ISO/IEC JTC1/SC32, Data Management and Interchange, January 2000, http://www.sdct.itl.nist.gov/~ftp/l8/sc32wg2/2000/events/openforum/index.htm

[6] Object Management Group. CORBA/IIOP 2.3.1 Specification. October 1999.

[7] W3C. Extensible Markup Language (XML), Version 1.0, http://www.w3.org/TR/REC-xml

[8] Data Entity Dictionary Specification Language (DEDSL) - Abstract Syntax, CCSDS 647.0-R-2.0, Draft Recommendation for Space Data System. Standards, Consultative Committee on Space Data Systems, November 1999.

[9] Special Issue: Planetary Data System, Planetary and Space Science, Pergamon, Volume 44, Number 1, January 1996.

[10] ISO/IEC11179-1,6, http://www.iso.ch/infoe/text.htm.

[11] NCI Common Data Elements Dictionary, http://cii-server5.nci.nih.gov:8080/pls/cde_public/cde_java.show.